

Summer Program FGV/EMAp 2019

# INTRODUCTION TO MACHINE LEARNING WITH PYTHON

Prof. Luis Gustavo Nonato

University of São Paulo - São Carlos - SP

# Course Structure and Content

# Course Structure and Content

- Principal Component Analysis

# Course Structure and Content

- Principal Component Analysis
- Regression

# Course Structure and Content

- Principal Component Analysis
- Regression
- Clustering

# Course Structure and Content

- Principal Component Analysis
- Regression
- Clustering
- Classification

# Course Structure and Content

- Principal Component Analysis
- Regression
- Clustering
- Classification
- Tree-based Regression and Classification

# Course Structure and Content

- Principal Component Analysis
- Regression
- Clustering
- Classification
- Tree-based Regression and Classification



ANALYTICS | BIG DATA | HADOOP | DATA PLUMBING | DATAVIZ | JOBS

The Algorithms Every Data Scientist Should Know

## regression

Ordinary Least Squares Regression (OLSR)  
Linear Regression  
Logistic Regression  
Multivariate Adaptive Regression Splines (MARS)  
Locally Estimated Scatterplot Smoothing (LOESS)  
Jackknife Regression

## regularization

Ridge Regression  
Least Absolute Shrinkage and Selection Operator (LASSO)  
Least-Angle Regression (LARS)

## instance based

also called case-based, memory-based

k-Nearest Neighbour (kNN)  
Learning Vector Quantization (LVQ)  
Self-Organizing Map (SOM)  
Locally Weighted Learning (LWL)

## dimensionality reduction

Principal Component Analysis (PCA)  
Principal Component Regression (PCR)  
Partial Least Squares Regression (PLSR)  
Genetic Mapping  
Multidimensional Scaling (MDS)  
Projection Pursuit  
Discriminant Analysis (LDA, MDA, QDA, FDA)

## deep learning

Deep Boltzmann Machine (DBM)  
Deep Belief Networks (DBN)  
Convolutional Neural Network (CNN)  
Stacked Auto-Encoders

## associated rule

Apriori  
Eclat  
FP-Growth

## ensemble

Single-Sample Boosting  
Bootstrapped Aggregation (Bagging)  
Adaboost  
Boosted-Sample-Selection Aggregation  
Gradient Boosting Machines (GBM)  
Gradient Boosted Regression Trees (GBRT)  
Random Forest

## think big data

## bayesian

Naive Bayes  
Multinomial Naive Bayes  
Averaged One-Dependence Estimators (AODE)  
Bayesian Network (BN)  
Bayesian Markov Models  
Bayesian Markov Models  
Conditional random fields (CRFs)

## decision tree

Classification and Regression Tree (CART)  
C4.5 and C5.0 (different versions of a powerful approach)  
C4.5-based Automatic Interaction Detection (CAID)  
Decision Stump  
Random Forests  
Conditional Decision Trees

## clustering

Single-linkage clustering  
k-Means  
k-Medians  
Expectation Maximization (EM)  
Hierarchical Clustering  
Fuzzy clustering  
DBSCAN  
OPTICS algorithm  
Non Negative Matrix Factorization  
Latent Dirichlet allocation (LDA)

## neural networks

Self Organizing Map  
Perceptron  
Back-Propagation  
Hopfield Network  
Radial Basis Function Network (RBFN)  
Backpropagation  
Autoencoders  
Hopfield networks  
Boltzmann machines  
Restricted Boltzmann Machines  
Spiking Neural Networks  
Learning Vector quantization (LVQ)

## ...and others

Support Vector Machines (SVM)  
Cross-validation  
Inductive Logic Programming (ILP)  
Reinforcement Learning (Q-Learning, Temporal Difference, State-Action-Reward-State-Action (SARSA))  
ANOVA  
Information Fuzzy Network (IFN)  
Page Rank  
Conditional Random Fields (CRF)



# Class Dynamics

Our lessons will be divided in two parts:

# Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content

# Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content
- 2 Lab: Practice with Python

# Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content
- 2 Lab: Practice with Python

Put ML algorithms to work in practice, using real data sets !!

# Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content
- 2 Lab: Practice with Python

Put ML algorithms to work in practice, using real data sets !!

No deep knowledge in Python will be required.

# Class Dynamics

Our lessons will be divided in two parts:

- 1 Theoretical content
- 2 Lab: Practice with Python

Put ML algorithms to work in practice, using real data sets !!

No deep knowledge in Python will be required.

The code will be as simple as possible and easily understandable, even for students not so experienced in computer programming.

# Machine Learning: What and Why?

# Machine Learning: What and Why?

*“Given an email, is it a spam?”*



# Machine Learning: What and Why?

*“Given an email, is it a spam?”*

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

# Machine Learning: What and Why?

*“Given an email, is it a spam?”*

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Data can make up for the lack of knowledge !!

# Machine Learning: What and Why?

*“Given an email, is it a spam?”*

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Data can make up for the lack of knowledge !!

It is assumed that there is a “hidden” model/process capable of analyzing the content of an email and then decides whether it is a spam or not.

# Machine Learning: What and Why?

*“Given an email, is it a spam?”*

We can hardly come up with an algorithm to answer this question without some knowledge about what characterize a spam.

Data can make up for the lack of knowledge !!

It is assumed that there is a “hidden” model/process capable of analyzing the content of an email and then decides whether it is a spam or not.

The goal is to *learn* the hidden model/process from data !!

# Taxonomy

There are a variety of learning methods.

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

## **1 Predictive or Supervised**

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

## **1 Predictive or Supervised**

- Regression:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$



# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

## 1 Predictive or Supervised

- Regression:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

## 1 Predictive or Supervised

- Regression:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given  $\mathbf{x}$  predict  $y$

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

## 1 Predictive or Supervised

- Regression:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given  $\mathbf{x}$  predict  $y$

## 2 Descriptive or Unsupervised

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

## 1 Predictive or Supervised

- Regression:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given  $\mathbf{x}$  predict  $y$

## 2 Descriptive or Unsupervised

- Clustering

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

## 1 Predictive or Supervised

- Regression:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given  $\mathbf{x}$  predict  $y$

## 2 Descriptive or Unsupervised

- Clustering
- Manifold Learning

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

## 1 Predictive or Supervised

- Regression:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given  $\mathbf{x}$  predict  $y$

## 2 Descriptive or Unsupervised

- Clustering
- Manifold Learning
- Relationship between instances

# Taxonomy

There are a variety of learning methods.

Existing methods can simplistically be divided into two categories:

## 1 Predictive or Supervised

- Regression:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$
- Classification:  $(\mathbf{x}, y), \mathbf{x} \in \mathbb{R}^d, y \in \{c_1, \dots, c_k\}$

Given  $\mathbf{x}$  predict  $y$

## 2 Descriptive or Unsupervised

- Clustering
- Manifold Learning
- Relationship between instances

The taxonomy above is not comprehensive, there are methods that do not properly fit in any of those two categories. For example, semi-supervised and reinforcement learning methods.

# Recap

Summary of the Lecture:



# Recap

## Summary of the Lecture:

- This course covers important concepts and machine learning techniques

# Recap

## Summary of the Lecture:

- This course covers important concepts and machine learning techniques
- There are, though, relevant not covered topics (Neural Networks for instance)

# Recap

## Summary of the Lecture:

- This course covers important concepts and machine learning techniques
- There are, though, relevant not covered topics (Neural Networks for instance)
- We will adopt a very practical approach, with real data and applications.